AUDIO-VISUAL SPEAKER RECOGNITION FOR VIDEO BROADCAST NEWS

Chalapathy V. Neti, Andrew Senior [*]
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598

1. SUMMARY

Significant progress has been made in the transcription of the audio stream in the broadcast news domain for both radio news and TV news (HUB4 task). Such transcripts provide an excellent means of indexing video content for search and retrieval. Speaker identification is an important technology in this domain both for selecting high-accuracy speaker-dependent models for transcription and as an index for search and retrieval of video content. However, the transcription accuracy under acoustically degraded conditions (such as background noise) and channel mismatch (telephone) still needs further improvements. To make improvements in such degraded conditions is a hard problem. We have begun investigating the combination of audio-based processing with visual processing for both speech and speaker recognition to improve the accuracy in acoustically degraded conditions. The use of two independent sources of information brings significantly increased robustness to signal degradation since degradations in the two channels are uncorrelated, and the use of visual information allows a much faster speaker identification than possible with acoustic information. In this paper, we present some encouraging preliminary results for audio-visual speaker recognition for TV broadcast news data (CNN).

2. METHOD

The system carries out speaker identification independently on the acoustic and visual signal. The results for the two modes are then combined together to arrive at a final speaker identity, and a list of scores for all the registered speakers indicating their similarity to the test speaker.

2.1. Visual speaker identification

The visual mode of speaker identification is implemented as a face recognition system. Faces are found and tracked in the video sequences, and recognized by comparison with a database of candidate face templates. This section describes the detection, tracking and recognition processes.

2.1.1. Face detection

Faces can occur at a variety of scales, locations and orientations in the video frames. In this system, we make the assumption that faces are close to the vertical, and that there is no face smaller than 66 pixels high. However to test for a face at all the remaining locations and scales, the system searches for a fixed size template in an image pyramid. The image pyramid is constructed by repeatedly downsampling the original image to give progressively lower resolution representations of the original frame. Within each of these sub images, we consider all square regions of the same size as our face template (typically 11x11 pixels) as candidate face locations. A sequence of tests is used to test whether a region contains a face or not. These are described in more detail in another paper [2].

First, the region must contain a high proportion of skin-tone pixels, and then the intensities of the candidate region are compared with a statistical face model. The statistical model returns a score-based on a linear discriminant classifier and 'Distance from face space' [4]. A high combined score from both these face detectors indicates that the candidate region is indeed a face. In the experiments described here, where a single face is present in each clip, it suffices to find the highest-scoring candidate region. Candidate face regions with small perturbations of scale, location and rotation are also tested and the maximum scoring candidate chosen, giving refined estimates of these three parameters.

In subsequent frames, the face is tracked by using a simple algorithm to predict the new face location, and using the statistical model to search for the face in candidate regions near the predicted location and with similar scales and rotations. A low score is interpreted as a failure of tracking, and the algorithm begins again with an exhaustive search.

[*] We would like to thank H. Beigi for providing acoustic speaker identification code

2.1.2. Face recognition

Having found the face, K facial features are located using the same technique used for face detection. As many as 29 facial features are used, including the hairline, chin, ears, and the corners of mouth, nose, eyes and eyebrows. Prior statistics are used to restrict the search area for each feature. At each of the estimated feature locations, a Gabor Jet representation [5] is generated. A Gabor jet is a set of 2-dimensional Gabor filters — each a sine wave modulated by a Gaussian. Each filter has scale (the sine wave frequency and Gaussian variance) and orientation (of the sine wave). We use five scales and eight orientations, giving 40 complex coefficients $(a_j, j = 1,..40)$ at each feature location.

A simple distance metric is used to compute the distance between the feature vectors for trained faces and the test candidates. The distance between the ith trained candidate and a test candidate for feature k is defined as:

$$\mathcal{S}_a^k(\mathcal{J},\mathcal{J}^i) = \frac{\sum_j a_j a_j^i}{\sqrt{\sum_j a_j^2 \sum_j (a_j^i)^2}} \tag{1}$$

A simple average of these similarities, $S_i = 1/K \sum_1^K \mathcal{S}_a^k(\mathcal{J},\mathcal{J}^i)$, gives an overall measure for the similarity of the test face to the face template in the database.

2.2. Audio-based speaker identification

The IBM Speaker identification system uses two techniques: a model-based approach and a frame-based approach [1]. In the experiments described here, we use the frame-based approach for speaker identification based on audio. Briefly, the frame-based approach can be described as follows:

Let M_i be the model corresponding to the i^th enrolled speaker. M_i is represented by a mixture Gaussian model defined by the parameter set $\{\mu_{i,j}, \Sigma_{i,j}, \gamma_{i,j}\}_{j=1,..n_i}$, consisting of the mean vector, covariance matrix and mixture weights for each of the n_i components of speaker i's model. These models are created using training data consisting of a sequence of M frames of speech with d-dimensional cepstral feature vectors, $\{f_m\}_{m=1,..,M}$. The goal of speaker identification is to find the model, M_i, that best explains the test data represented by a sequence of N frames, $\{f_n i\}_{n=1,..,N}$. We use the following frame-based weighted likelihood distance measure, $d_{i,n}$ in making the decision:

$$d_{i,n} = -\log[\sum_{j=1}^{n_i} \gamma_{i,j} p(f_n|\mu_{i,j}, \Sigma_{i,j}] \tag{2}$$

The total distance, D_i of model M_i from the test data is then taken to be the sum of the distances over all the test frames.

$$D_i = \sum_{n=1}^{N} d_{i,n} \tag{3}$$

2.3. Fusion

In general, mode-fusion or the integration different modes of information can be acheived by any of the following methods of data fusion [3].

- data fusion — this involves integration of different modalities in raw form e.g., video camera and microphone outputs.
- feature fusion — features are extracted from the raw data and subsequently combined. This involves e.g., audio features for speaker or speech recognition with visual features of the face for speaker recognition.
- decision fusion — this is the fusion at the most advanced stage of processing and involves combining the decisions of two different classifiers making independent decisions about the identity of the speaker-based on audio and visual features

In general, decision fusion provides a higher degree of robustness, but is accompanied by possible loss of information. An optimal fusion policy of using one of these fusion strategies or some weighted combination of the three strategies needs to be investigated. In this paper, we experimented with the technique of decision fusion and combine the decisions based on visual information (face-identification) and audio information (based on audio speaker identification).

Given the audio-based speaker recognition and face recognition scores, *audio-visual speaker identification* is carried out as follows: the top N scores are generated-based on both audio and video-based identification schemes. The two lists are combined by a weighted sum and the best-scoring candidate is chosen. The combined score $score_i^{av}$ is defined as:

$$score_i^{av} = \alpha * D_i + \beta * S_i \tag{4}$$

Since the two classifiers used for audio-based speaker identification and visual face identification are different, the mixture weights α and β are chosen manually such that the two scores after the multiplication are appropriately normalized. All the results presented use a single set of weights α and β, but making the weights dependent on the confidence of the acoustic classifier or the visual classifier would improve the results by biassing towards the more reliable decision.

Acoustic Condition	Audio Id	Video Id	audio-video Id
Clean	100%	70%	95%
Channel mismatch	40%	70%	65%
Noise mismatch	52%	70%	80%

Table 1. Audio-visual speaker ID t:efficiency

3. RESULTS

We collected 20–40 second clips of video data for anchors and reporters with frontal shots of their faces from the HUB4 video data. The training data contained 20 clips and the test data contained 20 additional clips of the same speakers.

Table 1 shows the results of experiments performed-based on a database of 20 training speakers and 20 test speakers from the MPEG2 encoded CNN video news data. Note that under clean conditions audio-based speaker identification is 100% accurate for the small dataset. Channel mismatch was simulated by using a telephone channel filter on the audio data. Noise mismatch was created by adding speech noise to the audio signal at a signal-to-noise ratio of about 10 dB. Note that under clean conditions (row 1 of the Table) audio-based speaker identification is 100% accurate for the small dataset. However, under acoustically degraded conditions (telephone channel and additive speech noise) it degrades to 40% and 52%, respectively. Combining this with video-based face identification with an accuracy of 70% improves the joint accuracy to 65% and 80% respectively.

The full benefit of the video accuracy is not realized when the audio channel is degraded by telephone channel mismatch. However, when the mismatch is due to additive noise, the joint recognition accuracy is better than either audio-based identification or video-based identification.

Analysis of the speakers for whom the joint audio-visual identification is better than either showed that the best candidates based on audio alone and video alone are different from each other. However, the true candidate was second or third best. Thus the score of the true candidate after combination is the best.

Figure 1 shows the results of joint audio-visual speaker identification under a variety of experimental conditions corresponding to different levels of audio-based speaker identification accuracy and video-based speaker identification accuracy. Note that in all cases, combining with video identification improves the audio-based accuracy, except in the clean audio conditions where the joint accuracy is slightly lower than the audio-based accuracy of 100%.

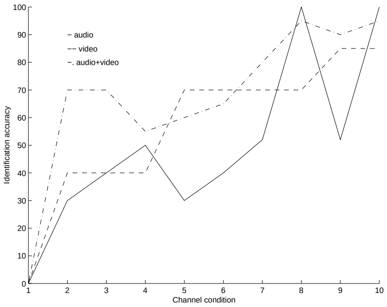

Figure 1. Identification accuracy vs experimental condition: solid line shows audio-only; dashed line for video only; dash-dot for audio+video

REFERENCES

[1] H. Beigi, S. H. Maes, U.V. Chaudari, J.S. Sorenson IBM Model-based and frame-by-frame speaker recognition. *Speaker Recognition and Its commercial and Forensic applications*, Avignon, France, 1998.

[2] A. Senior Face and Feature Finding for a Face recognition System. To appear in *Proceedings of the Audio and Vidoe-based Person Authentication'99*, March 1999.

[3] David L. Hall, Mathematical Techniques in multisensor data fusion. Artech House, 1992.

[4] M. Turk, A. Pentland Eigenfaces for Recognition. *Journal of Cognitive Neuro Science* Vol. 3 No. 1 pp. 71–86 1991.

[5] L. Wiskott, C. von der Malsburg Recognizing Faces by Dynamic Link Matching. *Proceedings of the International Conference on Artificial Neural Networks* pp. 347–352 1995.